DSO 428
Business Analytics Project Report (Final Deliverable)
Project Title: Film Market-Entry Analysis. 06 Dec 2019
Group 8

Movie Industry Context and Background:

A significant and non negligible part of current society, movies have not only contributed greatly to many economies with industry taxes and job opportunities but also encouraged cultural exchange and idea communication between individuals. With a global box office net worth reaching $41.7 billion, the movie industry has become a prime target for many investors. However, the field can be very complicated for newcomers. To make a movie, film investors need to consider numerous factors, ranging from the selection of movie production companies, genres and scripts to choosing the right management and creative teams (casts).

To help mitigate the risk of entering the market for this project's **target audience--movie investors and producers**--, the business analyst team Group 8 has prepared a revenue-focused and data-driven market research. Specifically, the team aims to provide not only **prediction model that estimates possible outcomes based on certain attributes chosen** but also **recommendations and actionable plans that can help movie investors achieve market success and high returns**.

Target Audience of the Analytics Project

As mentioned earlier, the target audience is movie investors and producers, mostly those who are new to the market. Answers for the core questions above will **inform them of the general industry status** (risks and benefits of entering the market), **specific attributes/factors** (such as actors, production companies or genres) **to pay attention to** and lastly **actionable plans and prediction to lead them to final movie success and mitigate the market entry risks**.

Dataset Overview

Data source:

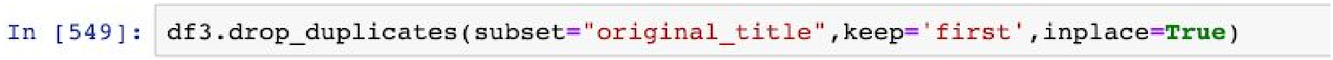
- Kaggle "The Movie Dataset"
- https://www.kaggle.com/rounakbanik/the-movies-dataset#ratings.csv

Tables: (Linked by movieId)

- movies_metadata.csv: The main Movies Metadata file contains information on 45,000 movies featured in the Full MovieLens dataset. Its features include posters, backdrops, budget, revenue, release dates, languages, production countries and companies.
- keywords.csv: Contains the movie plot keywords for our MovieLens movies.
- credits.csv: Consists of Cast and Crew Information for all our movies.
- links.csv: The file that contains the TMDB and IMDB IDs of all the movies featured in the Full MovieLens dataset.
- links_small.csv: Contains the TMDB and IMDB IDs of a small subset of 9,000 movies of the Full Dataset.
- ratings_small.csv: The subset of 100,000 ratings from 700 users on 9,000 movies.

The initial data cleaning process:

- The original Kaggle dataset is filtered using **Excel**. Specific procedures include:
  - Excluding blank or bogus titles
  - Filtering out entries with Budget < $100,000
    - This is because we found many movies with budget listed below 100,000 to be not true to their actual value. Thus, if we included these bogus values, the ROI calculated for them would also be incorrect, which would distort our analysis results.
  - Preserving data entries with only release dates after the year 1931
- Comments were added in code where columns were dropped because they were not used in a specific analysis. (eg. keyword analysis was run with a separate method using SQL and Excel)
- Duplicates are also dropped, using the codes shown by the screenshot below.



```
In [549]: df3.drop_duplicates(subset="original_title",keep='first',inplace=True)
```

Analysis Project Plan Overview and Core Questions to Answer:

To give a detailed, comprehensive analysis, the project is dissected into **three sub-sections**:

1. **Data Exploration**
   Core Question #1: What is the current market status of the movie industry?
   This part of the analysis will take a general look at the overall data to get a sense of the industry's risks and opportunities as well as some important factors impacting the movie production.

2. **Specific Attribute Analysis**
   Core Question #2: What are the determining factors for movies to gain high revenue/high ROI/market success?
   After learning about what factors can influence a movie's success from section 1, section 2 will examine in detail how each specific factor contributes to a movie's revenue and ROI.

3. **Prediction Model and Final Summary**
   Core Question #3: For a set certain budget, what is the recommended action plan in terms of detailed attributes such as the selection of actors, producers and keyword taglines?
   Combining the findings from both sections 1 and 2, a regression tree model is created to predict the success of a movie based on various attributes. A final summary with actionable recommendations will also be provided accordingly.

Tools and Software Used for the Analytics Project:

1. **Data exploration**
   - R, Python

   Good tools for data exploration, provide clear visualizations for descriptions of data and general trends

2. **Specific Attribute Analysis**
    - SQL
    - Tableau
    - Excel
    - Python

    Those are some of the industry standard tools for performing data analysis and we learned them in class

3. **Predictive Model**
    - Python (sklearn)

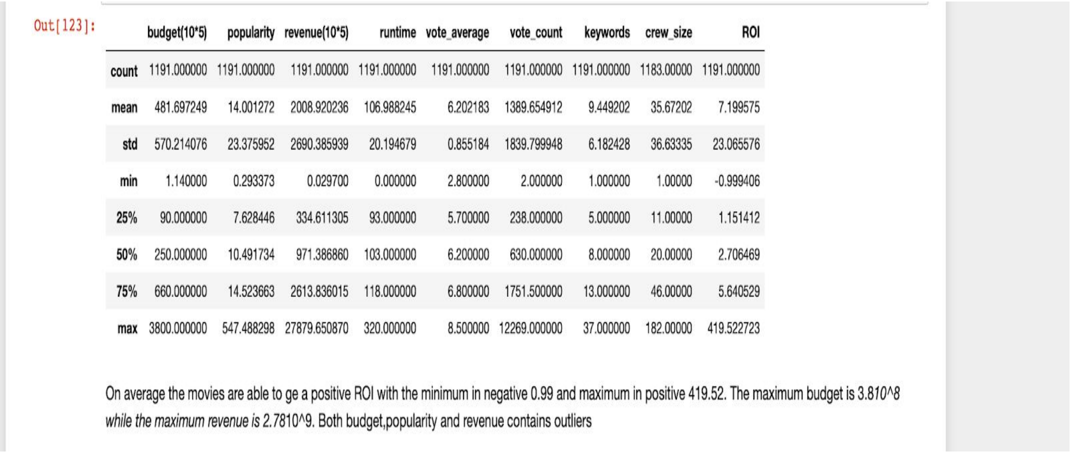
    Simple and powerful package for running machine learning model

Summary Statistics and Findings

PART I: Data Exploration and Overview

The first part of the analysis includes summary statistics or visualization of key metrics and characteristics of the overall movie market to provide incoming movie investors and producers a basic understanding of the data and the context.

Below is a general description of the dataset using Python on Jupyter notebook. A brief comment is also included in the screenshot under the table.

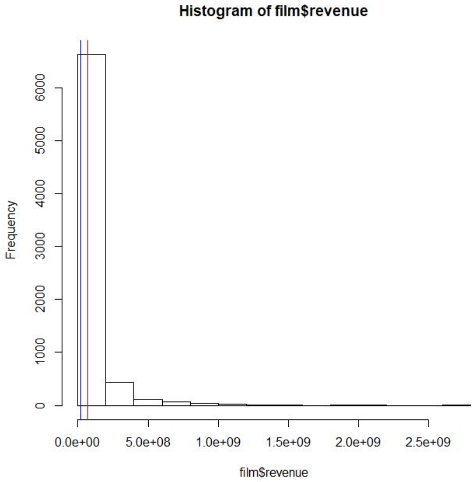


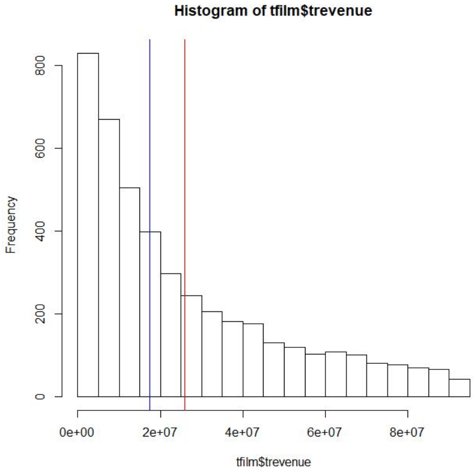
Out[123]:

	budget(10*5)	popularity	revenue(10*5)	runtime	vote_average	vote_count	keywords	crew_size	ROI
count	1191.000000	1191.000000	1191.000000	1191.000000	1191.000000	1191.000000	1191.000000	1183.00000	1191.000000
mean	481.697249	14.001272	2008.920236	106.988245	6.202183	1389.654912	9.449202	35.67202	7.199575
std	570.214076	23.375952	2690.385939	20.194679	0.855184	1839.799948	6.182428	36.63335	23.065576
min	1.140000	0.293373	0.029700	0.000000	2.800000	2.000000	1.000000	1.00000	-0.999406
25%	90.000000	7.628446	334.611305	93.000000	5.700000	238.000000	5.000000	11.00000	1.151412
50%	250.000000	10.491734	971.386860	103.000000	6.200000	630.000000	8.000000	20.00000	2.706469
75%	660.000000	14.523663	2613.836015	118.000000	6.800000	1751.500000	13.000000	46.00000	5.640529
max	3800.000000	547.488298	27879.650870	320.000000	8.500000	12269.000000	37.000000	182.00000	419.522723

On average the movies are able to ge a positive ROI with the minimum in negative 0.99 and maximum in positive 419.52. The maximum budget is 3.8*10^8 while the maximum revenue is* 2.7810^9. Both budget,popularity and revenue contains outliers

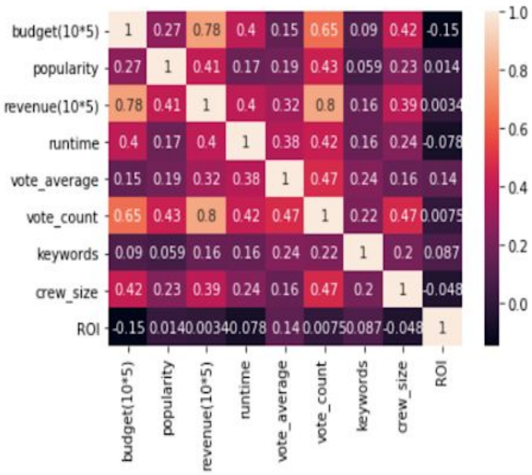
Next, we ran a correlation chart using python to find the relationship between each attribute

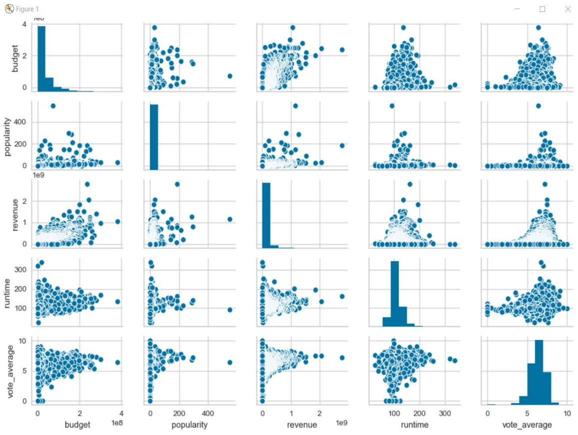




We used R to create two histograms showing the distribution of revenues from our dataset. The left histogram shows the data with all non-zero values, the histogram on the right shows the distribution of revenues **after 20% data trimming**. Applying this statistical tool helped us observe that the data is **skewed to the right, meaning that less and less movies are able to make higher revenues**. The vertical red line on the histogram represents the mean value, and the vertical blue line represents the median. What we found is that, while the median value remained almost the same after 20% data trim (from $17,381,034 to $17,381,942), the mean dropped significantly (from $69,766,838 to $25,823,720). It proves that there are outliers in our dataset, such as Toy Story movie with a revenue of over $2 billion. On the other hand, there are also movies with minimal revenues, making no profit.

As shown with the heatmap below, **budget and vote_count has a strong positive correlation with revenue.** However, those two variables are very difficult to measure before any movie is released, the deeper analysis will focus on factors that could contribute to those two variables such as actors/actresses, directors, and keywords.





PART 2: Specific Attribute Analysis (Deeper Analysis)

*Based on the initial general data analysis and findings, various impacting factors within the movie production process are discovered. Further analysis is conducted below to answer specific sub-questions developed under the core questions--**What are the determining factors for movies to gain high revenue/ high ROIs/ market success?** Domain knowledge of team members is applied to interpret further the initial results and dive deeper into the data for additional insights.*

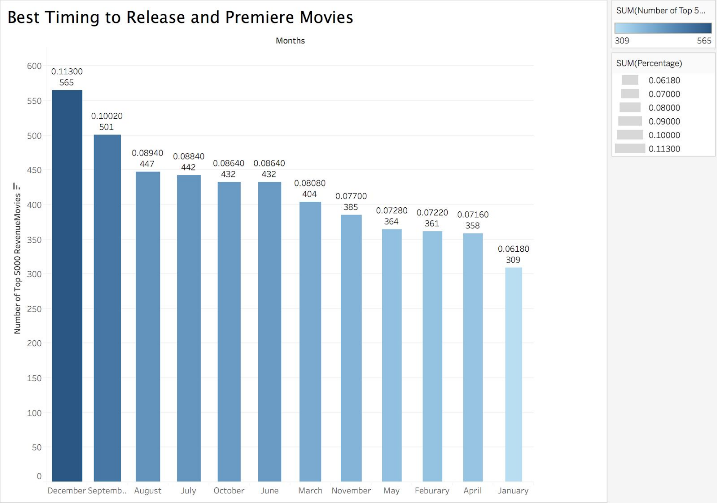
*Specifically, the following analysis is divided into **two sections--ROI and revenue--depending on whether the impacting factors involve budget input.** For example, the selection of tagline keyword or genre does not necessarily need budget investment. However, the decision of choosing which directors or production companies is highly dependent upon how much budget one has.*

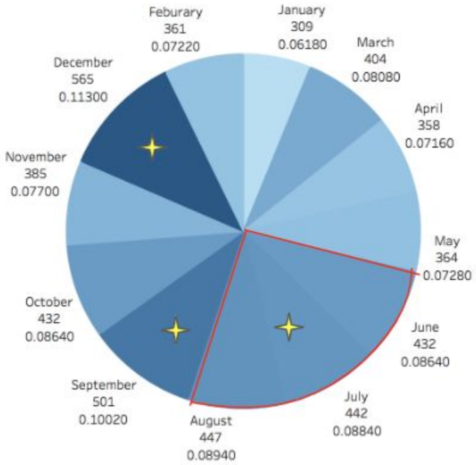
1. How certain factors contribute to a movie's final revenue [factors not heavily impacted by movie investors' budget amount]

a. Timing

Subquestion: What is the best timing to release/premiere movies for each genre?

i. For all movies

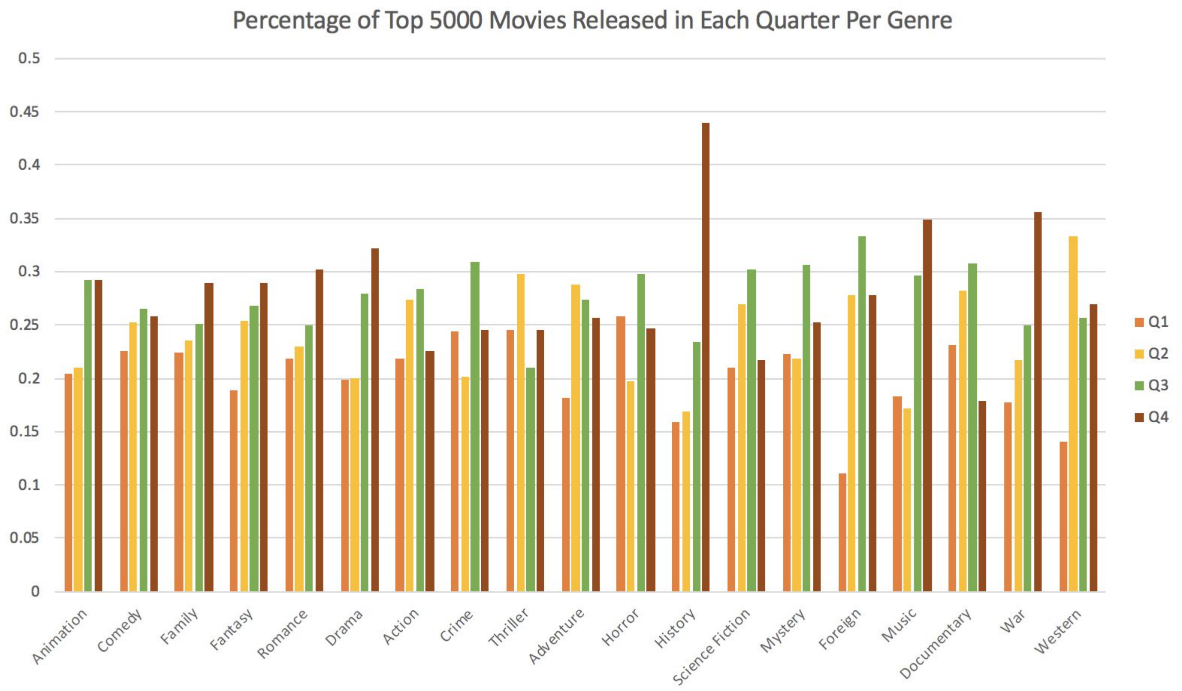




Using SQL and Excel, Top 5000 revenue movies (the top 10% out of 45000 data entries) with a valid budget are found and separated based on the months they were released in. The result is further visualized into a bar chart and pie chart using Tableau with color gradient scales and various sizes. Overall, the distribution is **relatively even** for each month. Notably, **December has the highest number of high revenue movie releases because it is during the holiday season when many families and friend groups are triggered to go to movies due to societal norms.**

Additionally, as seen from the pie chart, it is discovered that many high revenue movies are released in the **summer season (from June to August)** when most sports games are not on to compete with the movie market and many students are less occupied with school. Though **September** does not seem like a reasonable month to release movies as it is the start of sports and school season, many movies are also released during this time **because**

of film festivals like Toronto International and Venice where many potential Oscar contenders and independent films are released and distributed.

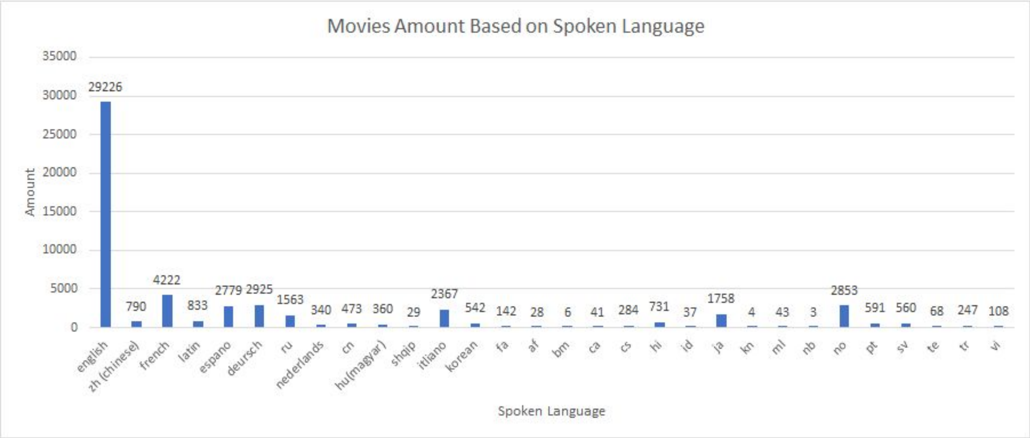


Moving onto timing by genre, similar to the general pattern, movies of most genres are released **relatively evenly for all four quarters**. However, certain **genres** are more **season-dependent**. For example, horror, crime and mystery movies are released in the third quarter as it is the summer scare/early Halloween season; on the other hand, movies of drama, romance and fantasy genres are mostly released in the last quarter possibly due to the holiday season.

Interestingly, **movies of history genre** are also mostly released in the last quarter. It is speculated that this period of time is amid award season where most Hollywood studios release serious topic films to be considered for Oscars and other major movie awards.

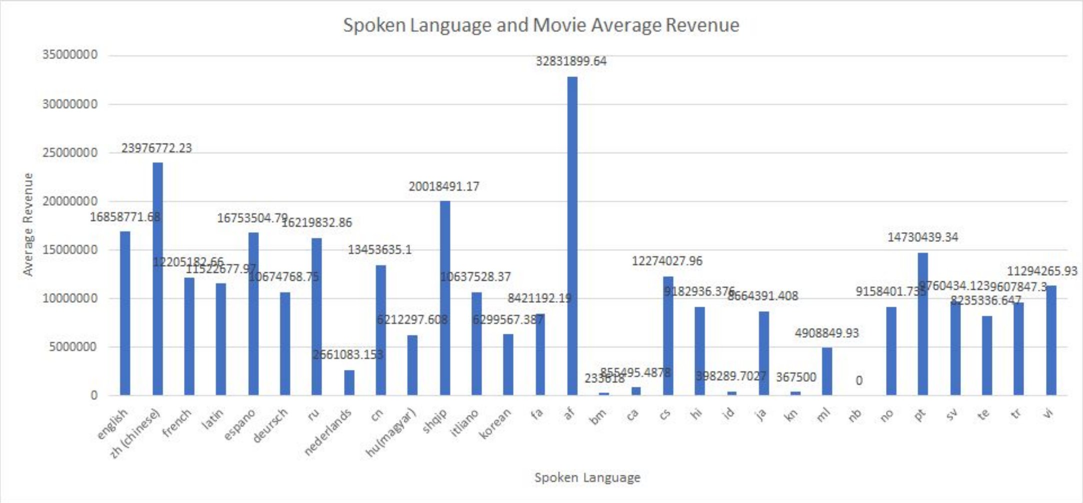
However, the data analysis reflects only the common situation. Realistically, a few movies released in “off-season” have also received high revenue. In conclusion, **movie investors should use this finding as a reference while deciding when to distribute their movies to confront/avoid competition and achieve the highest return possible.**

b. Language
Subquestion: Which language market yield the highest revenue
ii. Spoken Language and frequency used in movie dataset



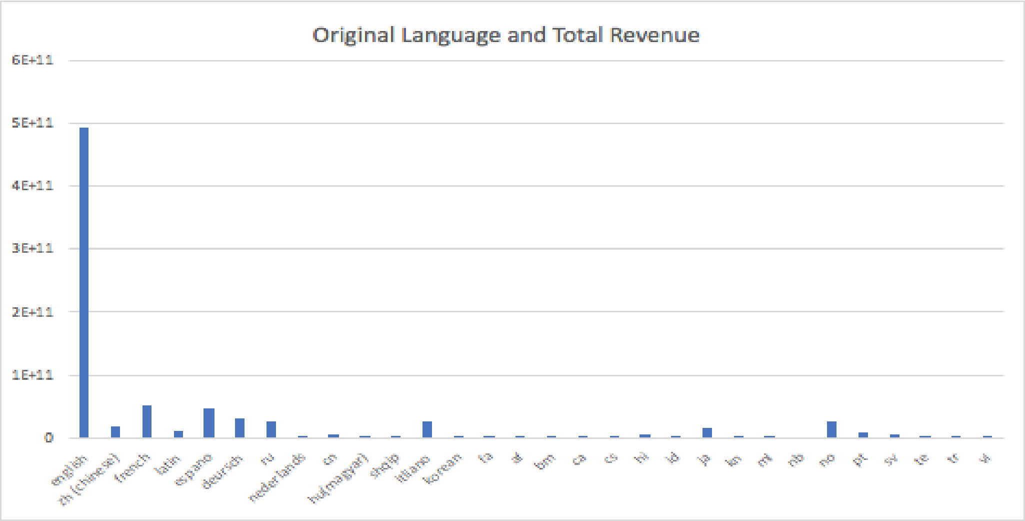
To run analysis over spoken language, first to clean up the dataset.Within this model, excel function(countif) is applied to get the frequency for each spoken language and how many movies in the dataset include that language. The result shows that English is the most frequently used spoken language, which is much higher compared to other spoken languages in the dataset.

iii. Spoken language and average revenue per movie



Within this model, in order to get the average movie revenue generated for each spoken language category, excel function (sumif) is used to get the total revenue by the amount of movies for each category of spoken language and divide the total revenue by the amount of movies for each category to get the average.

iv. Original Language and Spoken Language (why we choose original language over spoken language)



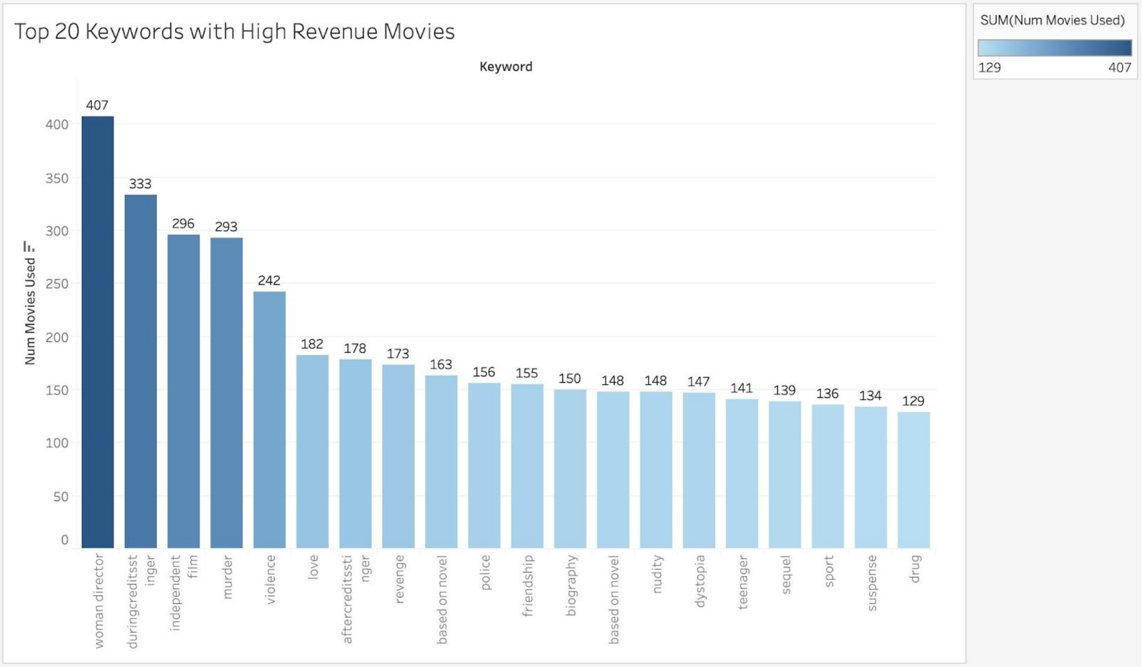
In order not to confuse audiences, the spoken language part was chosen to be discarded from the presentation. **While interpreting data models, English is the most frequently used spoken language and original language, but the result for the average revenue per movie turns out to be very different.** The result of spoken language and average revenue is not really convincing and not really helpful for the final conclusion.

For original revenue, it makes sense **because English and Chinese have the two largest speaking population groups.** The film industry differs from software, however, in functioning primarily as a cultural export. **English-language culture is the most globalized of all cultures, despite China having internal populations to rival the global English footprint.** For this reason, Hollywood is the dominant film industry in globalism and trade. **When a new movie comes out with the original language either in Chinese or English, it usually means it would have a larger potential market audience than the movie with original language of other languages.**

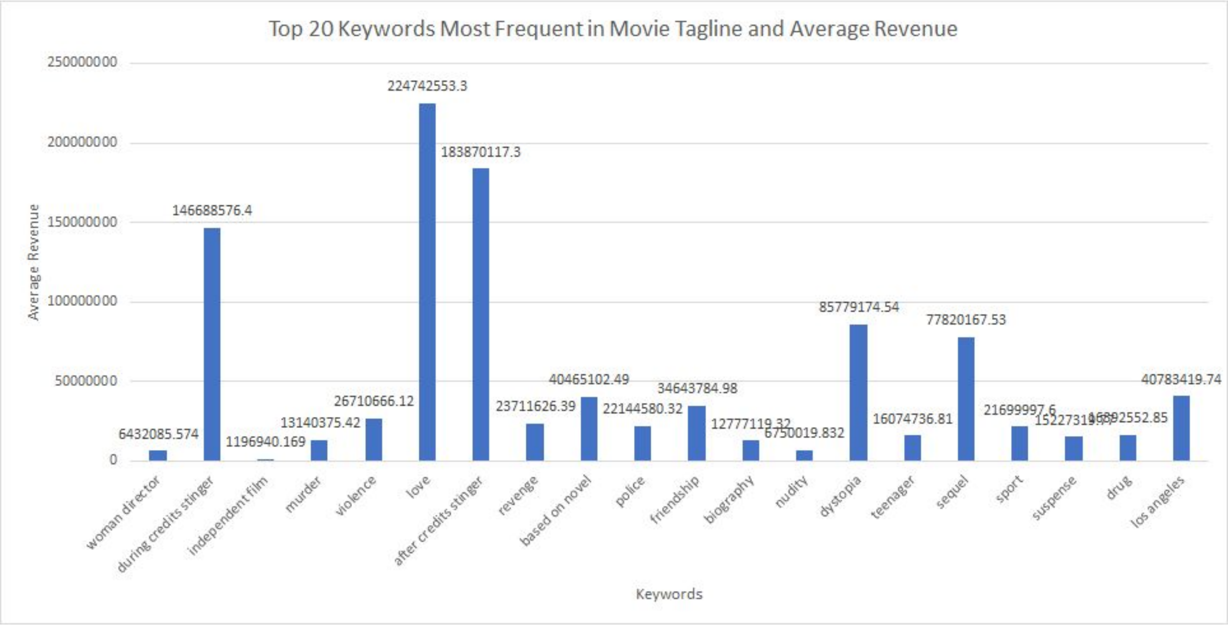
However, for spoken language, that is not the case. Movies that include low frequency spoken language tend to have a higher performance in revenue. This is because movies with higher chance of success (those with higher movie budget, famous producer company or belong to a successful series) **would tend to translate the movie to more spoken languages in order to open up the overseas market**. However, this information regarding spoken language and average revenue is **not helpful** to producer companies.

c. Keywords

Subquestion: What are the best keywords to use in tagline descriptions for high revenue movies?



Using SQL, the top 5000 movies with highest revenue are discovered. Afterward, the "Text to Column" function in Excel is used to separate and organized the casts data to one column of single-valued cells (shown as one attribute for further analysis). Lastly, SQL is used to find the frequency of high revenue movie appearance by keywords, which is further visualized via Tableau with scaled size and color.

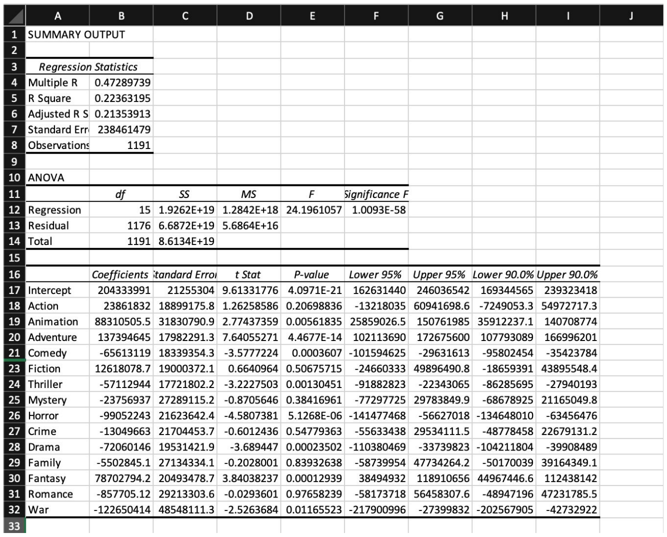


In order to get the top 20 keywords most frequently appear in movie tagline and the average revenue. Sumif function need be used to find the total revenue for those movies that use certain keywords in their taglines. Afterward, to get the average movie revenue it generates for each keyword category, the total revenue is divided by the amount of movies .

Interpretation: The keyword **“love” is the most frequently appeared keyword associated with highest average revenue**. As cliche as it may seem, customers still prefer **a splash of romance** in whatever movie they are watching, and movie production companies appears to be aware of that given the graph shown above.

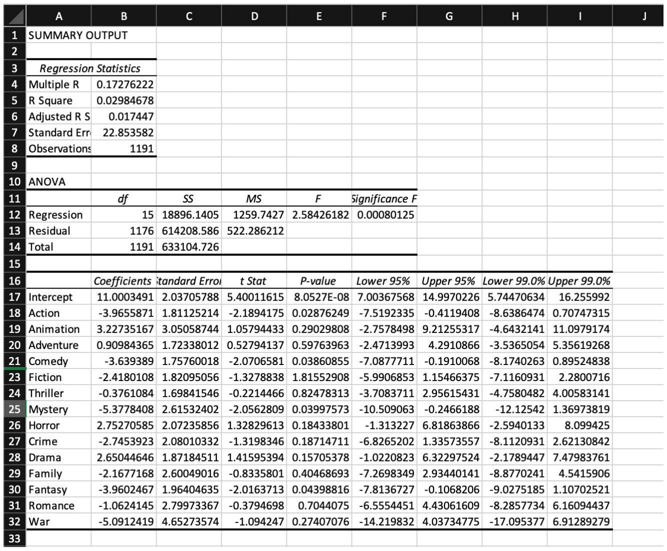
d. (1) Excel Linear Regression: Genre
Subquestion: Which movie genres matter more as a predictor for high revenue/ROI movies?

Against: Revenue || Return on Investment



	A	B	C	D	E	F	G	H	I	J
1	SUMMARY OUTPUT									
2										
3	*Regression Statistics*									
4	Multiple R	0.47289739								
5	R Square	0.22363195								
6	Adjusted R S	0.21353913								
7	Standard Err	238461479								
8	Observations	1191								
9										
10	ANOVA									
11		*df*	*SS*	*MS*	*F*	*Significance F*				
12	Regression	15	1.9262E+19	1.2842E+18	24.1961057	1.0093E-58				
13	Residual	1176	6.6872E+19	5.6864E+16						
14	Total	1191	8.6134E+19							
15										
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 90.0%*	*Upper 90.0%*	
17	Intercept	204333991	21255304	9.61331776	4.0971E-21	162631440	246036542	169344565	239323418	
18	Action	23861832	18899175.8	1.26258586	0.20698836	-13218035	60941698.6	-7249053.3	54972717.3	
19	Animation	88310505.5	31830790.9	2.77437359	0.00561835	25859026.5	150761985	35912237.1	140708774	
20	Adventure	137394645	17982291.3	7.64055271	4.4677E-14	102113690	172675600	107793089	166996201	
21	Comedy	-65613119	18339354.3	-3.5777224	0.0003607	-101594625	-29631613	-95802454	-35423784	
23	Fiction	12618078.7	19000372.1	0.6640964	0.50675715	-24660333	49896490.8	-18659391	43895548.4	
24	Thriller	-57112944	17721802.2	-3.2227503	0.00130451	-91882823	-22343065	-86285695	-27940193	
25	Mystery	-23756937	27289115.2	-0.8705646	0.38416961	-77297725	29783849.9	-68678925	21165049.8	
26	Horror	-99052243	21623642.4	-4.5807381	5.1268E-06	-141477468	-56627018	-134648010	-63456476	
27	Crime	-13049663	21704453.7	-0.6012436	0.54779363	-55633438	29534111.5	-48778458	22679131.2	
28	Drama	-72060146	19531421.9	-3.689447	0.00023502	-110380469	-33739823	-104211804	-39908489	
29	Family	-5502845.1	27134334.1	-0.2028001	0.83932638	-58739954	47734264.2	-50170039	39164349.1	
30	Fantasy	78702794.2	20493478.7	3.84038237	0.00012939	38494932	118910656	44967446.6	112438142	
31	Romance	-857705.12	29213303.6	-0.0293601	0.97658239	-58173718	56458307.6	-48947196	47231785.5	
32	War	-122650414	48548111.3	-2.5263684	0.01165523	-217900996	-27399832	-202567905	-42732922	
33										

^

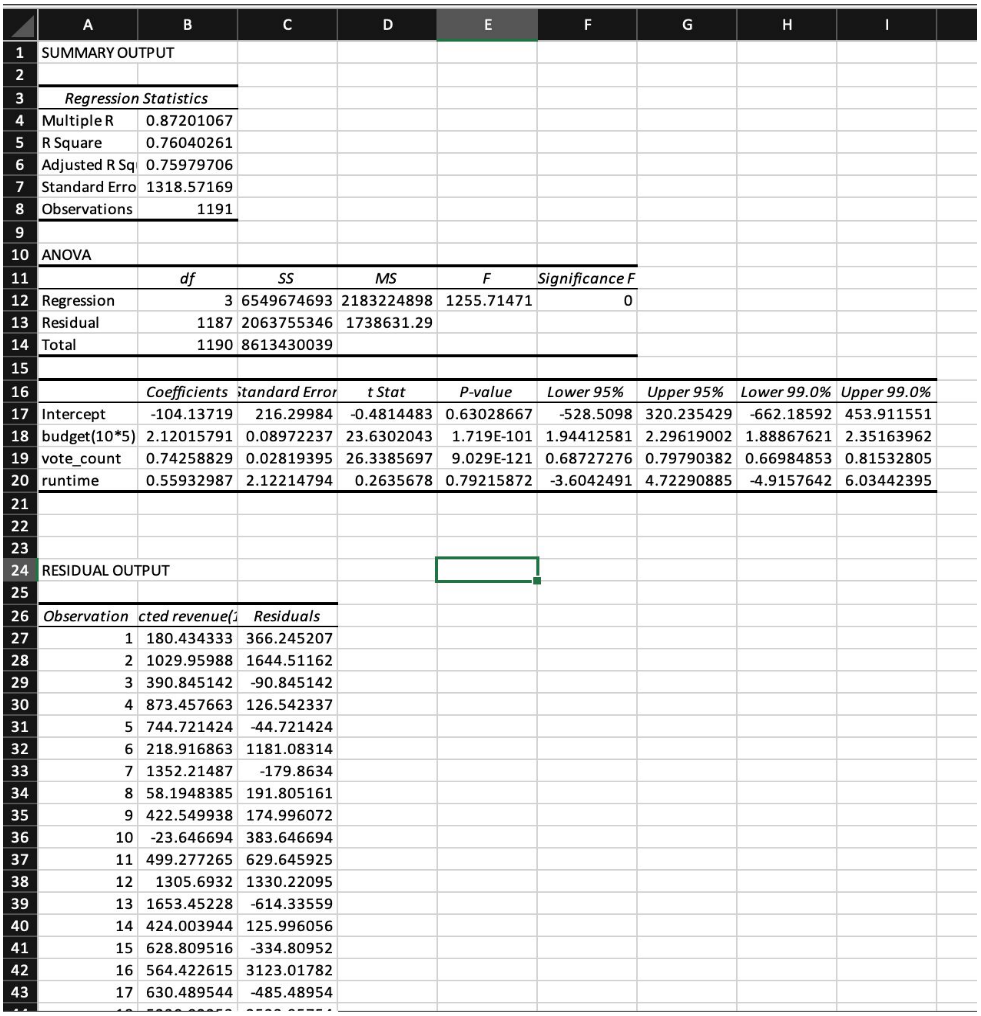


	A	B	C	D	E	F	G	H	I	J
1	SUMMARY OUTPUT									
2										
3	*Regression Statistics*									
4	Multiple R	0.17276222								
5	R Square	0.02984678								
6	Adjusted R S	0.017447								
7	Standard Err	22.853582								
8	Observations	1191								
9										
10	ANOVA									
11		*df*	*SS*	*MS*	*F*	*Significance F*				
12	Regression	15	18896.1405	1259.7427	2.58426182	0.00080125				
13	Residual	1176	614208.586	522.286212						
14	Total	1191	633104.726							
15										
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*	
17	Intercept	11.0003491	2.03705788	5.40011615	8.0527E-08	7.00367568	14.9970226	5.74470634	16.255992	
18	Action	-3.9655871	1.81125214	-2.1894175	0.02876249	-7.5192335	-0.4119408	-8.6386474	0.70747315	
19	Animation	3.22735167	3.05058744	1.05794433	0.29029808	-2.7578498	9.21255317	-4.6432141	11.0979174	
20	Adventure	0.90984365	1.72338012	0.52794137	0.59763963	-2.4713993	4.2910866	-3.5365054	5.35619268	
21	Comedy	-3.639389	1.75760018	-2.0706581	0.03860855	-7.0877711	-0.1910068	-8.1740263	0.89524838	
23	Fiction	-2.4180108	1.82095056	-1.3278838	1.81552908	-5.9906853	1.15466375	-7.1160931	2.2800716	
24	Thriller	-0.3761084	1.69841546	-0.2214466	0.82478313	-3.7083711	2.95615431	-4.7580482	4.00583141	
25	Mystery	-5.3778408	2.61532402	-2.0562809	0.03997573	-10.509063	-0.2466188	-12.12542	1.36973819	
26	Horror	2.75270585	2.07235856	1.32829613	0.18433801	-1.313227	6.81863866	-2.5940133	8.099425	
27	Crime	-2.7453923	2.08010332	-1.3198346	0.18714711	-6.8265202	1.33573557	-8.1120931	2.62130842	
28	Drama	2.65044646	1.87184511	1.41595394	0.15705378	-1.0220823	6.32297524	-2.1789447	7.47983761	
29	Family	-2.1677168	2.60049016	-0.8335801	0.40468693	-7.2698349	2.93440141	-8.8770241	4.5415906	
30	Fantasy	-3.9602467	1.96404635	-2.0163713	0.04398816	-7.8136727	-0.1068206	-9.0275185	1.10702521	
31	Romance	-1.0624145	2.79973367	-0.3794698	0.7044075	-6.5554451	4.43061609	-8.2857734	6.16094437	
32	War	-5.0912419	4.65273574	-1.094247	0.27407076	-14.219832	4.03734775	-17.095377	6.91289279	
33										

For this regression, dummy variables were used since genres do not possess numerical values. With reference to screenshots in part(d). The R-Sq shows a **22.4% correlation between the various genres and revenue, as opposed to 2.98% for genres vs ROI.** This is likely because **genres does not affect budget in a similar fashion to revenue which is influenced by customer preferences.** In other words, **budget and genre has little correlation.**

d. (2) Excel Regression:
Subquestion: We take variables that seem significant intuitively and perform regression on it to see if our claim(intuition) is/isn’t baseless.

- Dependent Variable: Revenue
- Independent Variables: budget, vote_count, runtime



	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.87201067							
5	R Square	0.76040261							
6	Adjusted R Sq	0.75979706							
7	Standard Erro	1318.57169							
8	Observations	1191							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	3	6549674693	2183224898	1255.71471	0			
13	Residual	1187	2063755346	1738631.29					
14	Total	1190	8613430039						
15									
16		*Coefficients*	*tandard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
17	Intercept	-104.13719	216.29984	-0.4814483	0.63028667	-528.5098	320.235429	-662.18592	453.911551
18	budget(10*5)	2.12015791	0.08972237	23.6302043	1.719E-101	1.94412581	2.29619002	1.88867621	2.35163962
19	vote_count	0.74258829	0.02819395	26.3385697	9.029E-121	0.68727276	0.79790382	0.66984853	0.81532805
20	runtime	0.55932987	2.12214794	0.2635678	0.79215872	-3.6042491	4.72290885	-4.9157642	6.03442395
21									
22									
23									
24	RESIDUAL OUTPUT								
25									
26	*Observation*	*cted revenue(1*	*Residuals*						
27	1	180.434333	366.245207						
28	2	1029.95988	1644.51162						
29	3	390.845142	-90.845142						
30	4	873.457663	126.542337						
31	5	744.721424	-44.721424						
32	6	218.916863	1181.08314						
33	7	1352.21487	-179.8634						
34	8	58.1948385	191.805161						
35	9	422.549938	174.996072						
36	10	-23.646694	383.646694						
37	11	499.277265	629.645925						
38	12	1305.6932	1330.22095						
39	13	1653.45228	-614.33559						
40	14	424.003944	125.996056						
41	15	628.809516	-334.80952						
42	16	564.422615	3123.01782						
43	17	630.489544	-485.48954						

As shown above, there is a **correlation of 76% between these 3 variables and revenue.** However, the **p-value of runtime is > 0.05, indicating that, at a 95% significance level, runtime is not a statistically significant predictor for revenue**, therefore it is excluded in our prediction model later on. Although vote_count does have a p-value < 0.05, it is still excluded from our prediction model because this information is not accessible pre-movie release.

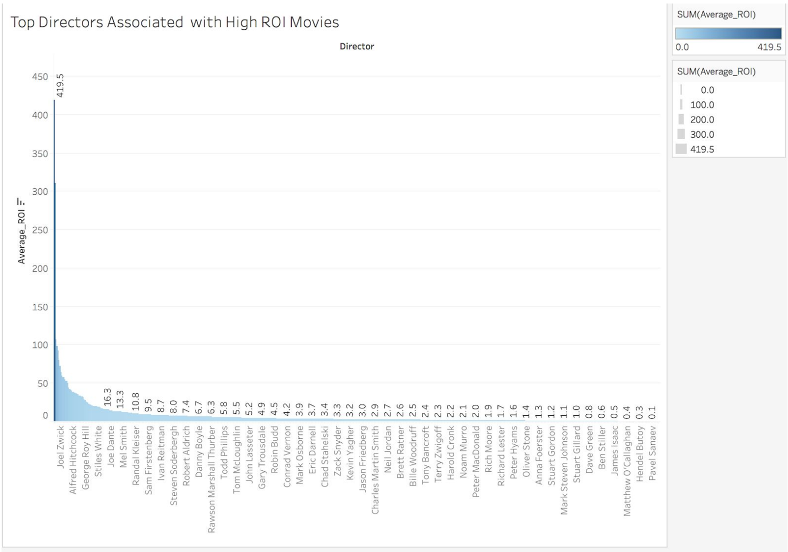
----

2. How certain factors contribute to a movie's ROI [factors impacted by movie investors' budget amount]

a. Directors

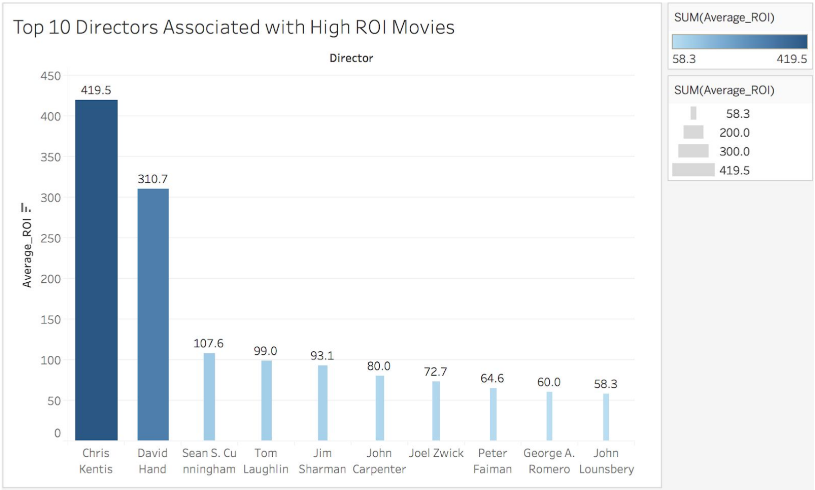
Subquestion: Which directors are associated with movies that yield high returns?

i. Overall pattern:



Using SQL and Excel, the top 1000 directors ordered by the average ROIs of their directed movies are discovered. Following that, results are visualized via Tableau with scaled size and color representing the amount of average ROIs. According to the graph pattern, it is shown that **a small number of directors are associated with a disproportionately large number of high ROI movies**. In other words, **for movie investors, there are only few director options and competition against other investors could be intense.**

ii. Specific Directors:

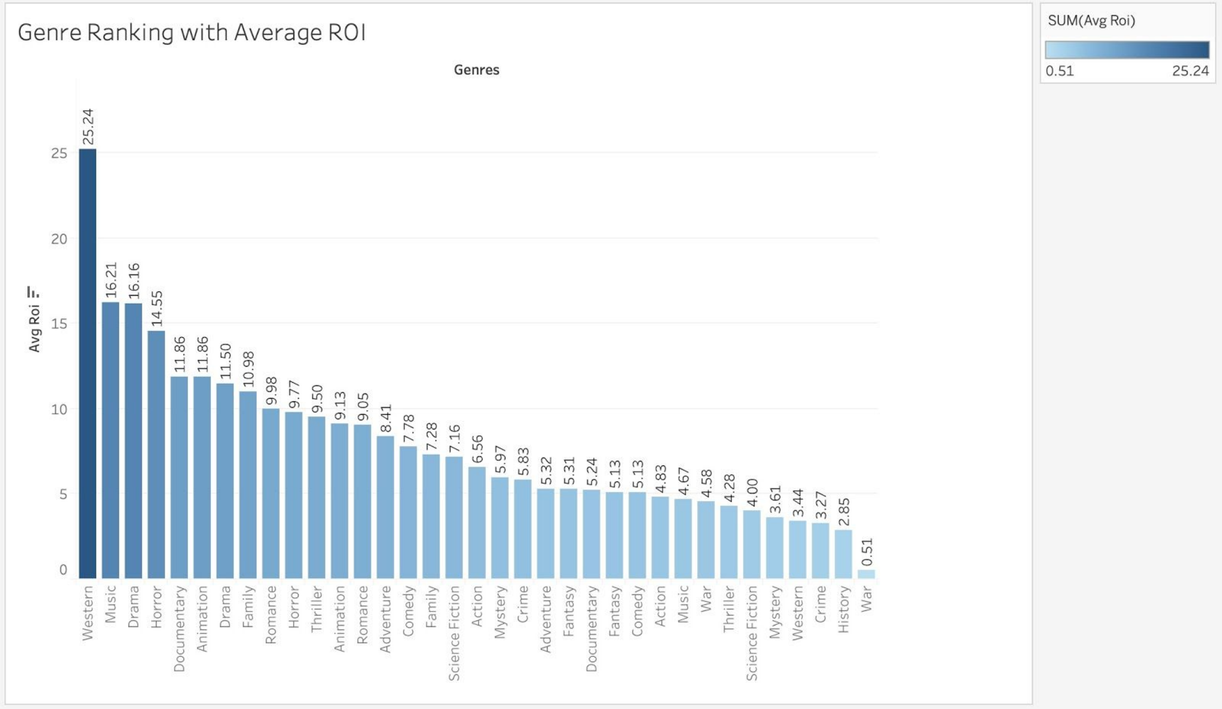


As shown, the top 10 directors with the highest average ROIs are displayed in the graph above. Corroborating with the overall excel linear regression analysis, which indicates that **thriller (cheap), animation and comedy genres** have high statistical significance on a movie's revenue success, most of these directors are associated with these genres. It is therefore further implied that, though movies of these genres vary a lot in revenues, they also have can achieve higher success if done right.

Interestingly, **most of the directors listed in the graph are not commonly seen or well-known by the public**. Notably, **Chris Kentis** becomes the director with the highest ROI (419.5) in the database by directing a **cheap thriller** “Silent House” with **only $6,000** over a time period of just four days. The movie eventually earned 12.8 million in revenue, which is not among the highest compared to other high revenue movies such as the “Harry Potter” series or “Avatar.” Like Kentis, other directors such as John Carpenter and Sean S. Cummingham also are known for **cheap thrillers**. According to this finding, it is suggested that, **for investors with lower budgets, cheap thrillers with one of these top directors could be a good market entry move.**

b. Genres

Subquestion: What are the main genres that high ROI movies belong to in recent years?



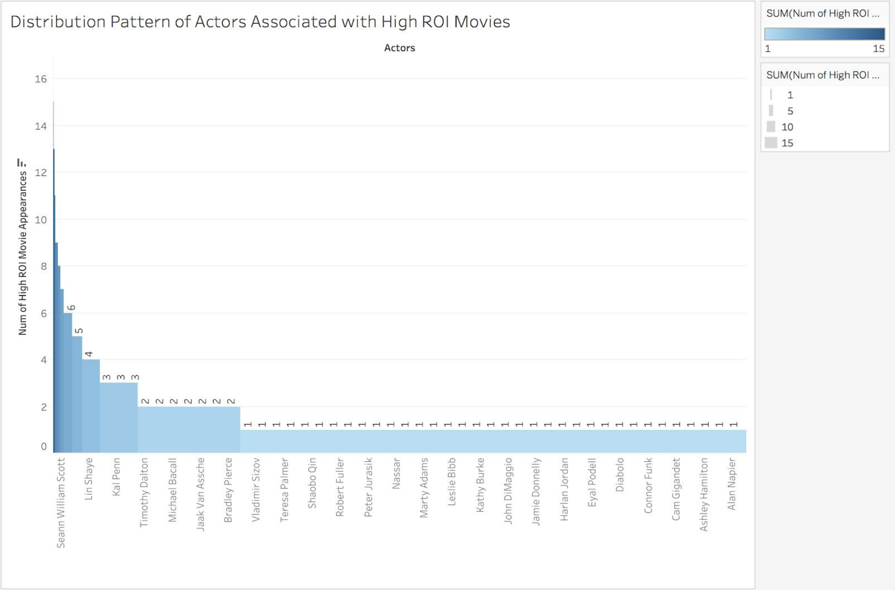
Through SQL, the top 1000 movies with highest ROIs are discovered by sorting the cleaned dataset with ROI in descending order. Afterward, the “Text to Column” function in Excel is used to separate and organized the genre data to one column of single-valued cells (shown as one attribute for further analysis). Finally, SQL query is conducted to find the average ROI of movies in each genre, which is further visualized via Tableau with scaled size and color.

According to the graph, its overall pattern shows that **ROIs vary greatly for different genres.** Notably, movies of **western, music, drama, horror and documentary genres have the highest average ROIs.** For movie investors who do not yet have a script or production plan, these genres could be promising to tap into.

c. Actors

Subquestion: What casts are hired/associated with high revenue movies?

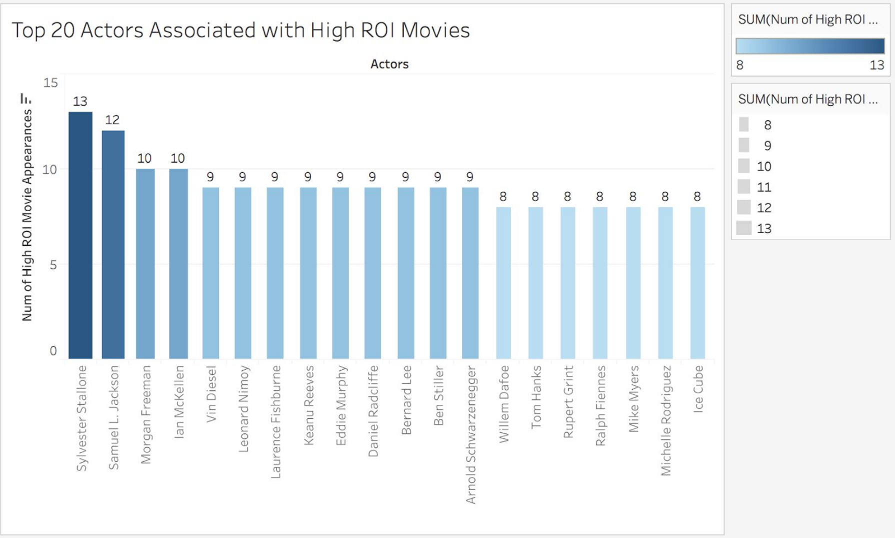
i. Overall pattern



Using SQL, the top 1000 movies with highest ROIs are discovered. Afterward, the "Text to Column" function in Excel is used to separate and organized the casts data to one column of single-valued cells (shown as one attribute for further analysis). Lastly, SQL is used to find the frequency of high ROI movie appearance by actors, which is further visualized via Tableau with scaled size and color.

As shown by the graph above, **most actors in the database are associated with only 1 or 2 high ROI movies**. Additionally, according to the graph's overall pattern, it is shown that, **similar to directors, a small number of actors are associated with a disproportionately large number of high ROI movies.** In other words, for movie investors, to recruit casts associated with high ROI movies, there are **only few options and competition against other investors could be intense.**

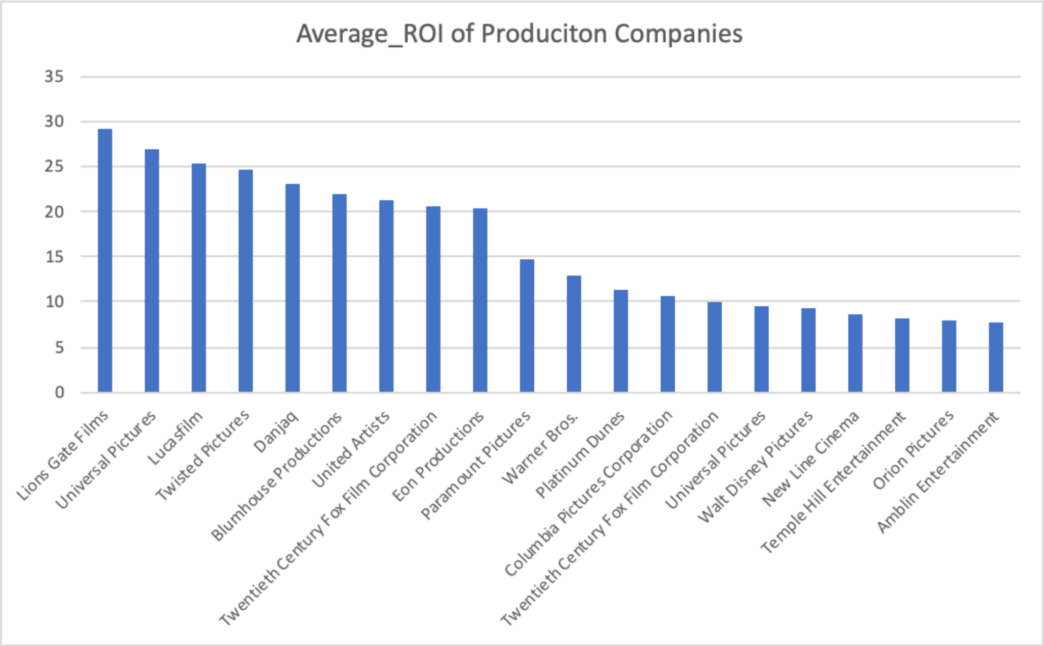
ii. Specific Actors:



The following graph contains the top 20 actors who have the highest frequency of appearance in high ROI movies according to our analysis. And as the graph indicated, top actors like **Sylvester Stallone, Samuel L. Jackson and Morgan Freeman** have appeared in over 10 high ROI movies, showing their **consistently good performance, versatility to tackle various genres of movies and therefore popularity and cost-efficiency as actors**. Unlike directors, result of this finding **corresponds to actors' reputation and large fan base, as most of these top actors are well-known among the public and have made frequent appearances** in high profile movies such as "Million Dollar Baby," "First Blood," and "Captain Marvel."

d. Production companies
Subquestion: Which production companies produces are able to profit more from their movies? Ie. Order production companies by their average ROI.

i. Top Twenty Production Companies with Highest Average ROI:

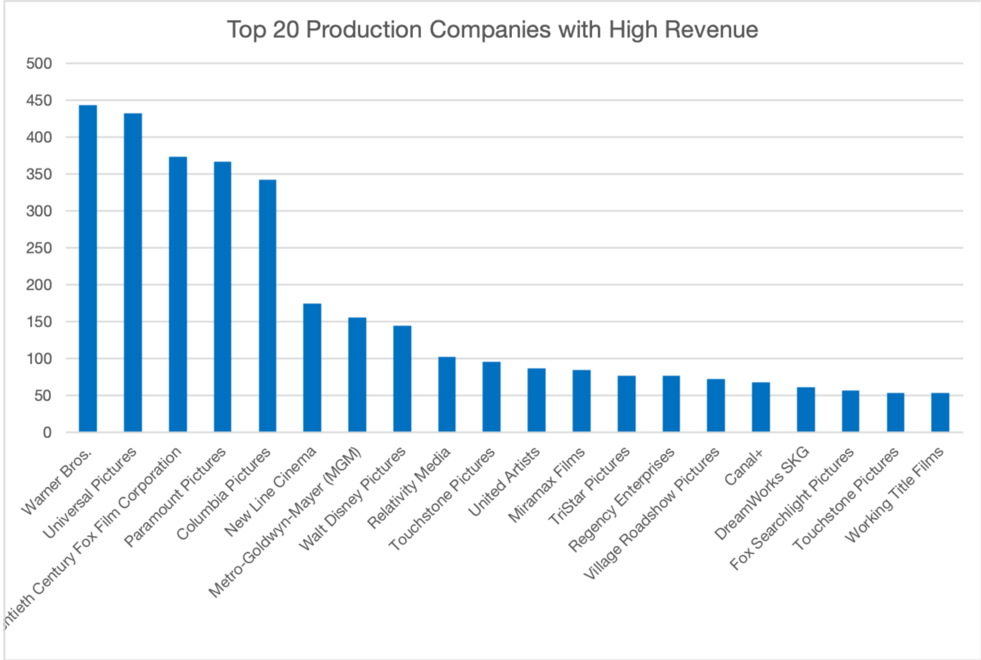


Using SQL and Excel, the top 20 production companies with the highest average ROI are discovered. The rows of movie data with both "Revenue" and "Budget" are retrieved before the analysis is performed. Then, the ROI formula (revenue-budget)/budget is used to get the ROI information. Since each movie has one or more production companies, the transformation of the data into two columns: company name and the ROI of all the movies they are involved with is required.

Next step, the spreadsheet is imported into Google BigQuery to run the analysis. The Average ROI of each production company is retrieved using the Group By function. A Having statement has to be incorporated in the query since some firms just have one movie produced which does not justify their capability to perform on the same level consistently. Therefore, setting a threshold to require producing more than 5 films in the dataset is necessary. (The reasons for setting the threshold to 5 are as below: taking into account the average number of films all the companies produced and trying to include the highest top revenue production companies in this top 20 ranking to get comprehensive understandings of their performance.) In the end, the top 20 companies' data is imported back to Excel for creating the above graph.

The result shows the top 20 companies which have high average ROI on the films they produced. This information is essential for investors when they are **identifying high return companies to invest their money in.** For example, compared to Warner Bros, Universal Pictures are having a much higher ROI in general, which means it is more attractive for investors. For some companies which have high revenue, it is not enough to determine whether they are good to invest in unless we know more information on their ROI. Therefore, analysis like this will provide a pretty good sense for investors.

ii. Top 20 Production Companies with High Revenue



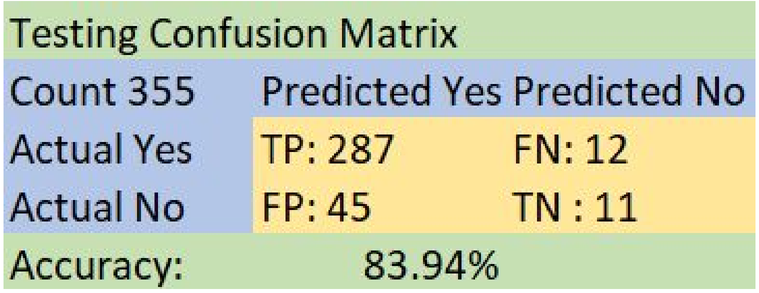
The graph shows a ranking of the average ROI of each production company. The results are different from running the analysis against revenue, in which results would be monopolized by large production companies. On the other hand, this result ranks production companies, **regardless of their scale(or budget)**, in terms of their **ability to generate more profit given their respective budgets**. In other words, **smaller production companies may earn less revenue, but have higher profit (percentage) than bigger production companies.**

Advanced Analytics (Data Mining, or Visualization/Dashboard, or Regression Analysis)

A **logistic regression model** is used to classify whether a movie will be successful or not. We define a movie as successful **if the movie's ROI is greater than the mean ROI of all movies** which is about 723%. A **70/30 split** was done for the training and testing purposes (training set contains 827 rows , testing set contains 355 rows).

Attributes included in the model: **Top 8 Genres by revenue, Top 20 production companies by ROI, Top 22 Actors/Actresses by ROI, Top 15 keywords by revenue,Top 10 director by Revenue and Top 10 director by ROI.**

The model was able to get an accuracy of **83.94%** on the testing dataset and a confusion matrix is provided below.



Testing Confusion Matrix		
Count 355	Predicted Yes	Predicted No
Actual Yes	TP: 287	FN: 12
Actual No	FP: 45	TN : 11
Accuracy:	83.94%	

Impact Statement & Summary

1. **Impact: To whom and how such findings or models would make a difference?**

   As the movie industry is a very competitive and high risk market, the model will allow the **investors** to significantly **reduce their risks of investing in a movie that will yield low ROI or even negative ROI**. The model could help movies investors to **predict** whether a movie will become successful **with the given attributes.** This will help investors avoid investing in movies that have a low chance of getting a high ROI. With an accuracy of **83.94%,** the investors should be fairly confident in the predicting power of the model.

   The predictive model will also be able to make recommendations on **what attributes the movie should get in order to increase the chance of being successful** (getting ROI higher than average ROI). Investors can test out what attributes or combination of attributes will help a movie improve its ROI. As many movies require a huge amount of budget, avoiding movies that could have low ROI would **save investors a lot of money**.

2. **Main take-aways**

   - All the analyses presented in this report are mainly based on movies with English as the language. Vote_average, Vote_count, popularity, are some of the variables excluded from the analyses as they are information only available after the movie is released, which is not the focus of Group 8's pre-release consultation to movie investors and producers that are new to the movie industry.

- Important findings from the analyses include **time of release, important keywords, actors/actresses that yield high ROI, directors that yield high ROI, the production companies with high ROI,** and **important genres**.

- The best times to release a movie would be in **December and Summer** season as there are more holidays, school breaks and less ongoing sports games and more people are inclined to watch movies.

- The keyword that appear frequently in top 5000 movies and have the highest average revenues is **love**.

- **English language movie** has the highest average revenue and Chinese language movie ranks at the second place after that.
- The top actors/actresses who have appeared more than 10 times in movies that yield high ROIs are **Sylvester Stallone, Samuel L. Jackson, Morgan Freedman, and Ian McKellen**.
-
- Besides that, many directors that yield high ROIs are less well-known ones such as **Chris Kentis, John Carpenter and Sean S. Cummingham**. Interestingly, all of their high ROI movies belong to the **cheap thriller** categories. For investors with **lower budgets,** cheap thrillers with one of these top directors could be a good market entry move.

- Besides well-known names such as **Lionsgate, Universal Pictures, Warner Bros, and Walt Disney**, top production companies with high ROI also include **a few companies of smaller scales and budget sizes** such as **Blumhouse and Orion**. For movie investors with less budget, these smaller production companies could be a good option.

- Last but not least, for investors, genres with highest average ROIs include **western, music, drama, horror and documentary**.

- Furthermore, after doing detailed analysis over various attributes, logistic regression is used to predict whether a movie will be successful. Investors could also test out what attributes or what combination of attributes within the attribute selection pool would yield **a better ROI**. with an accuracy of **83.94%**, the model uses the average ROI as the criteria to determine whether a movie is successful.

3. Summary:

In a nutshell, our consulting team are geared towards advice for movie investors and producers who are new to the market. As our clients provide specific properties for their upcoming film, we can use those information as input into our predictive model to **give a forecast or outlook on the movie's chances of success**. Given that, our clients might reconsider some of the film's properties and perform modifications that the team will also provide. For instance, depending on the positioning of the client's film, they might be advised to **change their partnering production company** from Walt Disney to Lionsgate, which is the most profitable production company based on our research. Then, **improvisations to the genre** of the film can also be made. An example of this would be, an originally 'Drama' labeled film may now be changed to include other genres such as **'Drama/Horror'** with slight changes to the script. Then, the consulting team may input these as variables in the predictive model to **recalculate the chances of success** for the newly modified film. If it results in a higher chance of success, the client would then be inclined to stick with that change, or continue making improvements for the highest chance of success possible within the timeframe before beginning production of the company.